Generating Low-Dimensional Denoised Representations of Nonlinear Data with Superparamagentic Agents

Thomas Ott†, Thomas Eggel† and Markus Christen‡

†Institute of Applied Simulation, ZHAW Zurich University of Applied Sciences
CH-8820 Waedenswil, Switzerland
‡University Research Priority Program Ethics, University of Zurich
CH-8008 Zurich, Switzerland
Email: christen@ethik.uzh.ch, eggl@zhaw.ch, ottt@zhaw.ch

Abstract—Visualisation of high-dimensional data by means of a low-dimensional embedding plays a key role in explorative data analysis. Classical approaches to dimensionality reduction, such as principal component analysis (PCA) and multidimensional scaling (MDS), struggle or even fail to reveal the relevant data characteristics when applied to noisy or nonlinear data structures. We present a novel approach for dimensionality reduction in combination with an automatic noise cleaning. By employing self-organising agents that are governed by the dynamics of the superparamagnetic clustering algorithm, the method is able to generate denoised low-dimensional embeddings for which the characteristics of nonlinear data structures are preserved or even emphasised. These properties are illustrated and compared to other approaches by means of toy and real-world examples.

1. Introduction

Classical approaches to dimensionality reduction aim to represent the data structure on a linear subspace of the original data space. For example, PCA performs a projection onto the axes with maximal data variance; whereas the goal of MDS is to find a low-dimensional embedding that preserves the interpoint distances. These methods often perform poorly when applied to nonlinear data structures. Furthermore, for many real-world applications, data vectors are not available. Instead, researchers are faced with similarity or proximity data with correct ordering, but potentially unreliable data values [1].

Various approaches exist to overcome these problems. The problem of possibly unreliable proximity data is addressed in non-metric MDS [2] by rescaling the proximities. To overcome the problem of nonlinearites, specialised nonlinear methods such as the Isomap algorithm [3] have been invented. Isomap requires constructing a k nearest neighbour graph to represent the structure of a data manifold. This enables a more correct description of proximities between points of a folded lower-dimensional manifold embedded into a higher-dimensional space, but there has been some debate about the distraction by noise.

In this contribution, we present a novel approach that is able to deal with nonlinear structures in data space and that includes a mechanism to reduce the distraction by noise. It takes advantage of a local, i.e., graph-based information akin to the Isomap approach and incorporates the non-metric MDS idea of applying a transformation to rescale the proximities. The core idea of the approach is to translate the data into a set of agents. These agents 'construct' a low-dimensional representation of the data in a self-organized way by moving according to laws of local spin interactions, as used for the superparamagnetic clustering algorithm [4, 5, 6]. In the following, we will describe the algorithm and illustrate its advantages in explorative data analysis using two toy examples and one real-world example. Latter uses data of an experiment on human similarity assessment of scientific disciplines that are used by the citation indexing service Web of Science.

2. Superparamagnetic Agents

We assume a given data set and its corresponding dissimilarity matrix with values $g_{ij} = g_{ji}$. Our method can be divided into two levels, where level 2 depends on level 1. In the first level, each data item is represented by a Potts spin variable and the dissimilarity matrix is encoded in the spin couplings. The spin system is treated in the formalism of the canonical ensemble, giving the probability for a certain spin configuration. One then can observe that the spins whose corresponding data items are similar tend to cluster in terms of the pair correlation G_{ij}, i.e., the probability of two spins being in the same state. By introducing a temperature-like parameter T, a cluster hierarchy can be generated. For smaller T, all spins tend to be in the same state. Upon an increase in T, large clusters break up into smaller clusters in a cascade of (pseudo-)phase transitions [4, 5]. For small T, spins that belong to data items of a noisy background can be filtered out as singletons that do not cluster.

In the second level, each data item is represented by an agent in a 2-dimensional coordinate system. The agents move according to laws that are governed by the local interactions of the spin system. In order to calculate G_{ij} a Markov chain Monte Carlo algorithm needs to be employed, which generates a sequence of binary pair correlation states $G_{ij}(t) \in \{0, 1\}$. Starting from a random distribution, two agents move towards each other if $G_{ij} = 1$, i.e., if the corresponding spins are in the same state in the current configuration, otherwise the agents drift apart, leading to a 2-dimensional distribution of agents.

Level 1: Spin system: Each Potts spin variable s_i can take possible values in $\{1, ..., q = 10\}$. Each spin is coupled to its k nearest neighbours (the choice of k is not critical, we choose $k = 10$ by default) and the couplings between spins are determined according to

$$J_{ij} = J_{ji} = \frac{1}{k} \exp\left(\frac{-g_{ij}^2}{2a^2}\right) \tag{1}$$

a is the average distance between neighbours. Each spin configuration s is associated via the Boltzmann distribution with the probability

$$p(s) = \frac{1}{Z} \exp(-H(s)/T) \tag{2}$$

with the Hamiltonian $H(s) = \sum_{(i,j)} J_{ij}(1 - \delta_{s_i s_j})$ and the normalization constant Z . The parameter T represents the system temperature. At a given T, the pair correlation $G_{ij} = \sum_s p(s)\delta_{s_i s_j}$ is calculated. G_{ij} is approximately calculated by

$$G_{ij} = \frac{1}{M} \sum_{t=1}^{M} \underbrace{\delta_{s_i^t s_j^t}}_{G_{ij}(t)} \tag{3}$$

where the Swendsen-Wang algorithm [7] has been used to generate the series of states.

Level 2: Agent system: A $\mathbb{R}^2-$embedding of a $n \times n$ matrix g_{ij} by means of superparamagnetic agents is constructed by setting up the superparamagnetic clustering framework and performing the following steps:

1. Choose a random agent distribution $(\vec{x_1^0}, ..., \vec{x_n^0})$ with $\vec{x_i^0} \in \mathbb{R}^2$
2. Choose a random spin configuration s^0
3. Set the temperature $T = T_{min}$ and ΔT
4. For T, calculate a new spin configuration s^{t+1} (according to Swendsen-Wang)
5. Calculate the actual pair correlations $G_{ij}(t+1) = \delta_{s_i^{t+1} s_j^{t+1}}$
6. For each pair of agents, do:
   - If $G_{ij}(t + 1) = 1$ and $J_{ij} > 0$ then

$$\vec{x_i^{t+1}} = \vec{x_i^t} + \alpha \cdot (\vec{x_j^t} - \vec{x_i^t}) \tag{4}$$

$$\vec{x_j^{t+1}} = \vec{x_j^t} + \alpha \cdot (\vec{x_i^t} - \vec{x_j^t}) \tag{5}$$

   - else

$$\vec{x_i^{t+1}} = \vec{x_i^t} + \beta \cdot e^{-d_{ij}^t} \cdot (\vec{x_i^t} - \vec{x_j^t}) \tag{6}$$

$$\vec{x_j^{t+1}} = \vec{x_j^t} + \beta \cdot e^{-d_{ij}^t} \cdot (\vec{x_j^t} - \vec{x_i^t}) \tag{7}$$

   where $d_{ij}^t = |\vec{x_i^t} - \vec{x_j^t}|$.

7. Set $T = T + \Delta T$ and go back to 4 as long as $T < T_{max}$
8. Agents whose spins are in no clusters even for T_{min} are removed (optional noise cleaning).

The choice of parameters is as follows:

* For the temperature range $[T_{min}, T_{max}]$ the optimal choice is the superparamagnetic phase since it provides information about the cluster structures. This range can differ for each data set, but the differences are usually small. Per default, we chose a fixed range of $[0, 0.1]$ and adapt it if necessary.

* ΔT is related to the number of Monte Carlo steps M: $\Delta T = [T_{min}, T_{max}]/M$. According to our experience, $M = 150$ gives stable results.

* $0 < \alpha < 0.5$ controls the attraction (speed) of two points whose spins are correlated.

* $0 < \beta$ controls the repulsion (speed) of two points whose spins are uncorrelated.

* The method does not offer unique solutions, which highlights the importance of the parameters involved. Simulations show that α and β strongly determine the scaling of the final agent configuration. α mainly affects the intra-cluster distances and β mainly affects the inter-cluster distances. For the examples in this paper we used the values $\alpha = 0.1$ and $\beta = 0.01$ that have proven useful to balance inter-and intra-cluster distances.

* The multiplier $e^{-d_{ij}^t}$ makes sure that the point configuration remains bounded.

3. Examples

3.1. Toy Set 1

In [8], a benchmark data set was introduced, showing two interlocked rings on a noisy background (750 points in total, 250 points for each ring and

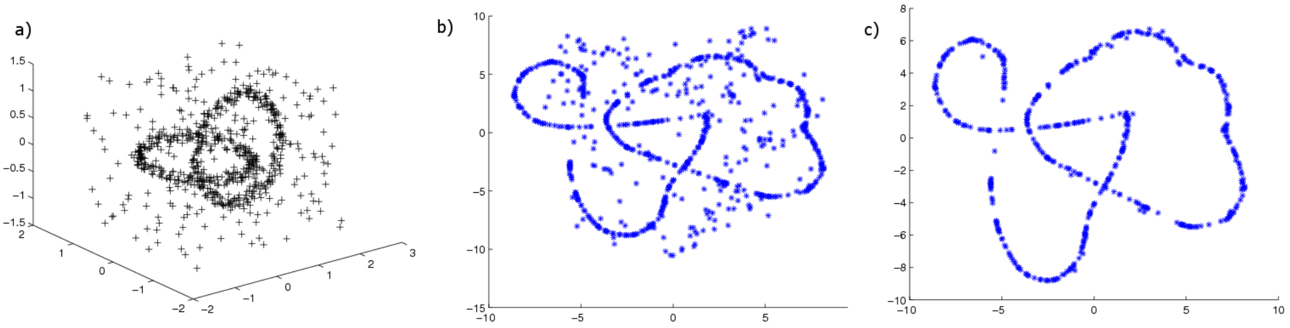

Figure 1: a) Original data set with two rings on a noisy background b) Superparamagnetic agent embedding without noise cleaning and c) with noise cleaning

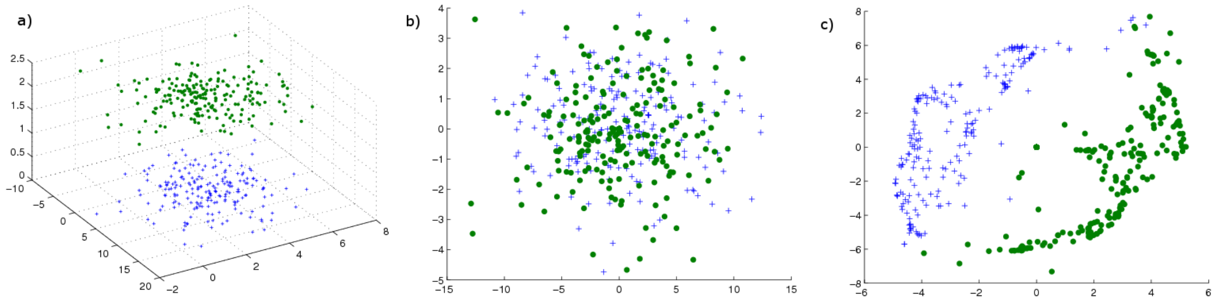

Figure 2: a) Original data set with two clusters b) PCA solution c) Superparamagnetic agent map (SAM)

the background, see Fig.1 a)). This problem cannot be solved by the majority clustering algorithms [8]. However, our approach is capable of generating a meaningful 2-dimensional image with inherent noise cleaning (Fig.1 c). Remind that the scaling in this image does not directly reflect the scaling in the original data. For example, the loop in one of the rings is a consequence of the dimensionality reduction.

3.2. Toy Set 2

The data set of this example consists of two Gaussian clusters with 200 points each and means $\mu_1 = (0,0,0)$ and $\mu_2 = (0,0,2)$ (Fig 2 a). The standard deviations are $\sigma_1^x = \sigma_2^x = 4.5$, $\sigma_1^y = \sigma_2^y = 1.5$ and $\sigma_1^z = \sigma_2^z = 0.05$. While the two clusters can clearly be distinguished in 3D, they are invisible to PCA in 2D because the extension in the $x-$ and $y-$direction is larger than in $z-$direction (Fig 2 b). For the superparamagnetic agents, this is no problem and a meaningful representation is generated (Fig 2 c).

3.3. Real World Example

We use data from a survey on the similarity of 249 scientific disciplines represented as subject categories that classify journals contained in the citation indexing and search service Web of Science provided by Thomson Reuters (http://thomsonreuters.com/thomson-reuters-web-of-science/). In the internet survey, the participants were presented with subject category X (including short descriptive text) as well as two other categories Y and Z and they had to choose to which discipline X is more similar. 876 researchers from all disciplines have been approached in multiple ways (e.g., via scientific associations) and they provided 33'558 assessments of the similarity of such subject category triplets, leading to a similarity matrix. To manage combinatorial explosion, we presuppose that disciplines from the same main fields (engineering, humanities, medicine, (social) science) are considered to be more similar when compared to a discipline from another field; i.e. participants that relate themselves to a specific field obtain random triplets where 90% emerge from "their" field. The task is robust for sequence effects and allows that subjects can stop the survey whenever they like. As similarity measure we use the ratio of positive attributions of two disciplines X and Y compared to the total number of possibilities to attribute X with Y.

Fig 3 shows the result. Although both approaches display a similar cluster discernibility, the topology of the original space is less well preserved in MDS compared to SAM. This is exemplified as follows: For

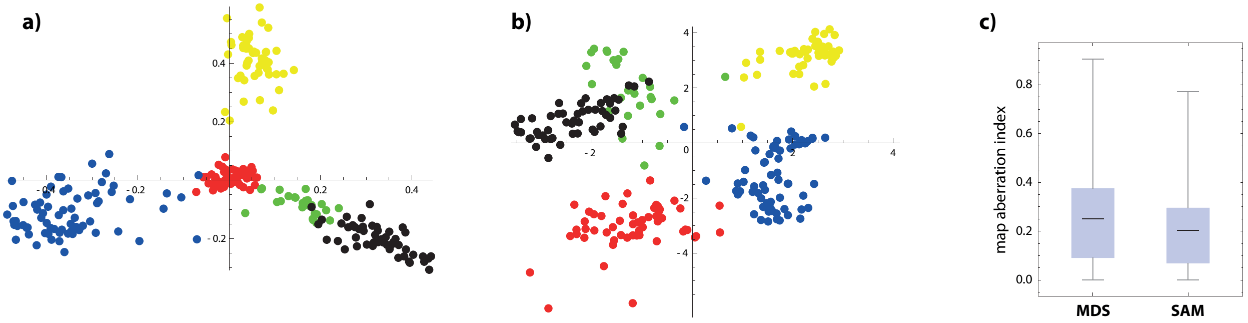

Figure 3: a) MDA solution (red: engineering, green: humanities, yellow: medicine, blue: science, black: social science b) Superparamagnetic agents map (SAM) c) Comparing map quality for humanities disciplines

each data item x_i, we calculate its distance to all other items $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ in the original space and in the map space and we normalize with the longest distance $\max\{d(x_i, x_j)\}$. We calculate for each item the sum of the absolutes of the normalized distance differences for each pair $\sum_j |\bar{d}_{orig.}(x_i, x_j) - \bar{d}_{map}(x_i, x_j)|$. The smaller the mean of this distribution (map aberration index), the better does the map preserve the topology of the original space. We show this for the group "humanities" for which most data was achieved in the survey (Fig 3 c).

4. Conclusion

We have introduced a novel algorithm for finding low-dimensional embedded representations of data described by a dissimilarity matrix g_{ij}, called superparamagnetic agents maps. The algorithm is based on a heuristic using superparamagnetic clustering. The main idea is that clustering provides the possibility to incorporate crucial information about cluster structures in the original data. Using this information, our superparamagnetic agents generate a low-dimensional image of the data. This approach has three advantages. First, it is capable of displaying nonlinear structures that are invisible to classical techniques such as PCA. Second, due to its robust nonparametric characteristics, it is able to distinguish between clusters and background noise. Third, it is superior in preserving the topology of the original data space. On the downside, the procedure is more time-consuming than other methods since it involves a spin system simulation.

Although the heuristic superparamagnetic agents algorithm was successful in several applications, questions remain regarding the theoretical understanding: How can we quantify the role of the parameters α and β? How can the theoretical connections to other methods such as nonmetric multidimensional scaling be elaborated? Can we also use the technique to determine the true dimensionality of higher-dimensional data structures? What other rules or clustering methods could be used instead of our heuristics to generate a low-dimensional representation? Answers would clarify why the heuristic superparamagnetic agents work so well in practice.

References

[1] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, S. Belongie, "Generalized Non-metric Multidimensional Scaling," *AISTATS, San Juan, Puerto Rico* (2007)

[2] J. B. Kruskal, "Nonmetric Multidimensional Scaling: A Numerical Method," *Psychometrika* 29, 115-129 (1964)

[3] J. B. Tenenbaum, V. de Silva, J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science 290* , 2319-2323 (2000)

[4] M. Blatt, S. Wiseman, E. Domany, "Superparamagnetic Clustering of Data," *Phys.Rev.Lett.* 76, 3251–3254 (1996)

[5] T. Ott, A. Kern, A. Schuffenhauer, M. Popov, P. Acklin, E. Jacoby, R. Stoop, "Sequential Superparamagnetic Clustering for Unbiased Classification of High-dimensional Chemical Data," *Journal of Chemical Information and Computer Sciences* 44(4), 1358-1364 (2004)

[6] T. Ott, A. Kern, W.-H. Steeb, R. Stoop, "Sequential Clustering: Tracking Down the Most Natural Clusters," *Journal of Statistical Mechanics: theory and experiment,* P11014 (2005)

[7] R. H. Swendsen, S. Wang, "Non-universal Critical Dynamics in Monte Carlo Simulations," *Phys.Rev.Lett* 58, 586-88 (1987)

[8] F. Landis, T. Ott, R. Stoop, "Hebbian Self-organizing Integrate-and-Fire Networks for Data Clustering," *Neural Computation* 22, 273-288 (2010)